Auditory and Visual Characteristics of Individual Talkers in Multimodal Speech Perception

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation
with distinction in Speech and Hearing Science
in the undergraduate colleges of The Ohio State University

by

Corinne D. Anderson

The Ohio State University
June 2007

Project Advisor: Dr. Janet M. Weisenberger,
Department of Speech and Hearing Science

Abstract

When people think about understanding speech, they primarily think about perceiving speech auditorily (via hearing); however, there are actually two key components to speech perception: auditory and visual. Speech perception is a multimodal process; i.e., combining more than one sense, involving the integration of auditory information and visual cues. Visual cues can supplement missing auditory information; for example, when auditory information is compromised, such as in noisy environments, seeing a talker's face can help a listener understand speech. Interestingly, auditory and visual integration occurs all of the time, even when the auditory and visual signals are perfectly intelligible. The role that visual cues play in speech perception is evidenced in a phenomenon known as the McGurk effect, which demonstrates how auditory and visual cues are integrated (McGurk and MacDonald, 1976).

Previous studies of audiovisual speech perception suggest that there are several factors affecting auditory and visual integration. One factor is characteristics of the auditory and visual signals; i.e., how much information is necessary in each signal for listeners to optimally integrate auditory and visual cues. A second factor is the auditory and visual characteristics of individual talkers; e.g., visible cues such as mouth opening or acoustic cues such as speech clarity, that might facilitate integration. A third factor is characteristics of the individual listener; such as central auditory or visual abilities, that might facilitate greater or lesser degrees of integration (Grant and Seitz, 1998).

The present study focused on the second factor, looking at both auditory and visual talker characteristics and their effect on auditory and visual integration of listeners.

Preliminary results of this study show considerable variability across talkers in the auditory only condition, suggesting that different talkers have different degrees of auditory intelligibility. Interestingly, there were also substantial differences in the amount of audiovisual integration produced by different talkers that were not highly correlated with auditory intelligibility, suggesting talkers who have optimal auditory intelligibility are not the same talkers that facilitate optimal audiovisual integration.

Acknowledgements

I would like to thank my advisor, Dr. Janet M. Weisenberger, for allowing me the opportunity to work with her on this research project and offering her constant guidance and support in the completion of this thesis. I would like to thank Natalie Feleppelle for all of the time and effort she contributed to ensure the success of this project through her personal and professional guidance. In addition, I would like to thank my family and friends for their constant encouragement during this process.

This project was supported by and ASC Undergraduate Research Scholarship and by the SBS Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

When people think about speech perception, they primarily think about perceiving speech auditorily, though there are two key components to speech perception: auditory and visual. Speech perception is a multimodal process involving the integration of auditory cues and visual cues. Integration occurs at all times, when the auditory and visual signals are compromised, such as in noisy or dark environments, as well as when the signals are perfectly intelligible. The occurrence of integration of visual and auditory cues is evidenced in a phenomenon known as the McGurk effect. McGurk and MacDonald (1976) demonstrated the integration of the two modalities by presenting observers with an auditory syllable, such as the bilabial “ba,” while simultaneously presenting observers with incongruent visual stimulus, such as the velar “ga,” which resulted in observers reporting to have perceived “da,” a fusion of the two places of articulation. This study shows that visual information is used despite the presence of an unambiguous auditory stimulus (McGurk and MacDonald, 1976).

Auditory and Visual Cues for Speech Perception

When producing speech, people use both vocal and visual signals to relay their message. However, when we think about speech, we generally think about it as an auditory process, perhaps because there are instances when we perceive speech adequately without the presence of visual cues, such as when we listen to the radio, or talk on the phone. Conversely, there are not as many instances requiring speech perception solely, if not primarily, using visual cues, as the task proves more difficult.

Through spectral and temporal aspects of the speech waveform, the auditory element of the speech signal conveys important information about place of articulation, manner of articulation, and voicing. The place of articulation refers to where the sound is produced, the manner of articulation refers to how the sound is produced, and the voicing refers to whether or not the vocal folds are vibrating. Through the visual signal alone, much less information is conveyed; the place and manner of articulation are often ambiguous and voicing is entirely indecipherable. With less information being relayed through the visual signal, some speech units may become indistinguishable, such as [b], [p] and [m]. Speech units such as these that have similar visual signals are known as *visemes* (Jackson, 1988).

While the auditory aspects of the signal play a more crucial role in speech perception, visual cues prove beneficial in enhancing speech perception. For example, when the signal is impaired, such as in hearing loss, one approach to improving speech recognition is to use speechreading, which relies heavily on visual stimuli through viseme recognition. However, visemes only allow speechreaders to distinguish between groups of sounds, such as with the viseme group /p, b, m/, which consists of bilabial stops. To distinguish between the units within the set, some auditory signal is necessary. For example, [p] is unvoiced and [b] and [m] are voiced, which cannot be distinguished through visual input alone (Jackson, 1988).

Viseme categories are not just determined by visible characteristics of the speech signal. Other factors that play substantial roles in visual speech perception and contribute to the organization of visemes include differences in articulation patterns among talkers and the environment in which they are produced. According to Jackson,

as speech visibility varies as a function of talker differences, no single viseme system exists across talkers. Across talkers, the number of viseme categories varies, as well as the components of the categories. Talkers who are easy to speechread exhibit more viseme categories than talkers who are less intelligible. Furthermore, "universal" viseme groups are more prominent among talkers who are easy to speechread (Jackson, 1988).

Within visemes, the term "homophenous," coined by Nitchie (see Jackson, 1988), refers to speech sounds or words that appear alike on the lips and cannot be distinguished by visual cues alone. So according to his classification system, though speech sounds within a viseme category may differ in voicing and/or nasality, they share the same place of articulation, such as /b/ and /p/, and are thus homophones (Jackson, 1988).

Audio-Visual Integration

Previous studies of audiovisual speech perception suggest that there are four factors affecting the integration process. The first factor is characteristics of the auditory and visual signals, particularly whether they are compromised in some way. The second factor is auditory and visual characteristics of the individual talker. The third factor is the type of task that requires speech perception. The final factor is characteristics of the individual listener. For each of these factors, one important question is whether redundancy of the auditory and visual stimulus is necessary for integration, or whether some ambiguity in the auditory or visual input results in optimal integration.

Characteristics of the Auditory and Visual Signals

The robustness of the auditory speech signal has been evidenced in a number of studies by measuring speech perception when the speech signal has been degraded in some form. A study by Remez et al. (1981) showed that the reduction of the speech waveform to three sine waves representing the first, second, and third formants in the original signal yields signals with sufficient information to support speech perception for both sentences and isolated syllables. These results show that there exists a degree of redundancy in the auditory signal, and that extensive removal of information from the auditory signal can still result in good intelligibility.

Researchers have extended the examination of the effects of degrading auditory signals to include effects of adding visual stimuli. Summerfield (1987) hypothesized that the addition of visual cues to the auditory signal would result in improved speech perception. He suggested that visual cues could aid in speech perception through the visual stimulus being redundant with auditory stimulus, emphasizing aspects of the signal and serving as a reinforcer. The second function that visual cues may serve is to complement the auditory cues, filling in information where the auditory signal may be lacking. The third role that visual cues may have is to generate temporal coincidence between the auditory and visual signals to highlight the most significant characteristics of the speech signal.

Grant and Braida (1991) continued to examine the effects of adding visual stimuli to a degraded auditory signal by determining the degree by which perception is improved. The study noted that adding visual cues to auditory speech in noise can

improve the signal-to-noise ratio by up to 15 dB (Sumby and Pollack, 1954), with each decibel increase resulting in an intelligibility improvement of 5-10 percent (Grant and Braida (1991).

A study conducted by Munhall et al. (2004) investigated the nature of the visual image processing during audiovisual speech perception by manipulating the spatial resolution of facial images for a speech-in-noise task. They found that speech signals were most commonly identified with the presentation of the unfiltered visual stimuli while least commonly identified with the presentation of the auditory stimulus only.

Auditory and Visual Characteristics of the Individual Talker

Jackson's study focused on talker characteristics, highlighting the negative effects of visual features, such as facial hair, thin or thick lips, etc. on their intelligibility. Individuals also vary in their auditory articulation, with some talkers' speech being more intelligible than others.

As noted above, it is known that some talkers have better speech intelligibility than others and that there is a large variance in the articulation used by talkers. When in an environment that makes communication difficult, many talkers engage in "clear speech" to improve intelligibility (Chen, 1980; Picheny et al., 1985; Uchanski et al., 1992; Peyton et al., 1994). When compared to conversational speech, a number of researchers have found clear speech to be marked by a slower speaking rate, increased temporal modulation, greater range of voice fundamental frequency, expanded vowel space, and more stimulus energy in high frequencies.

Gagne et. al. (2002) further investigated the use of clear speech by focusing on its benefits in auditory, visual, and audiovisual presentations. They suggested that articulation in clear speech production aids in perception of auditory, visual, and audiovisual conditions. Though results supported their theory, not all talkers produced clear speech benefits for all conditions. Additionally, the amount of benefit varied across talkers and across iterations within individual talkers. The question is whether a "good" talker is one that provides more benefit auditorily and/or visually or if a good talker is one that provides more ambiguity to allow for greater integration.

Tasks Requiring Speech Perception

Some researchers have argued that integration varies according to the specific task at hand. A study done by Grant and Seitz (1998) measured integration using a variety of auditory and visual materials, including isolated speech segments and sentences, having congruent and incongruent properties. The study related the different measures of integration ability to auditory-visual sentence benefit. The study reported substantial variability across subjects in auditory-visual integration for both sentences and nonsense syllables. A follow-up study done by Grant and Seitz (2000) showed that not only is there significant variability across listeners in the use of sentence context to facilitate word recognition, there are also considerable individual differences depending on the order in which isolated words or words in sentences are presented.

Individual Listener Characteristics

In the study mentioned above, Grant and Seitz (1998) examined whether individual listeners integrate auditory and visual cues with varying degrees of efficiency and found that there are significant levels of difference between listeners. One difference was that older subjects tended to be less efficient at integration than younger subjects. In a study done by Clark (2005), there was variability in the degree to which subjects exhibited the McGurk effect. When subjects were viewing themselves as talkers, half of the subjects showed a reduced McGurk effect, though none of the subjects exhibited particularly strong McGurk effects to these talkers.

The present study focused on talker characteristics, exploring aspects of individual talkers, both auditory and visual, and how they facilitate integration. The study specifically observed whether the talkers who produced good auditory perception alone or visual perception alone were the same talkers who produced good integration, which would help reveal what talker characteristics produce the best integration. Because the auditory signal carries more information than the visual signal, the talkers' speech was auditorily degraded by effectively reducing its' redundancy to avoid ceiling effects. Degrading the speech samples involved reducing it to three total sinusoids: F0+F1+F2. Digital video recordings of the speech were made for 14 different talkers. Participants were presented with these recordings and asked to identify the speech sound they perceived. We looked at integration performance of the participants under three conditions; 1) degraded auditory stimulus only 2) visual stimulus only, and 3)

degraded auditory + visual stimuli. The third condition contained both congruent and incongruent stimuli.

Chapter 2: Method

Participants

Participants in this study included 14 talkers and 10 observers. Of the talkers, there were seven males and seven females, all college or graduate students, ages 19-25. Of the observers, there were five males and five females, all college students, ages 19-22. All participants reported normal vision and were tested to have normal hearing capabilities. All are native English Speakers, with one observer being bilingual. One of the ten observing participants had completed undergraduate courses in phonetics, while the other nine participants had not taken courses containing information regarding phonetics and language. Observers received $80.00 for their involvement in this study.

Interfaces for Stimulus Presentation

Presentation of degraded auditory and visual stimuli was similar for all participants. Each participant was tested with stimuli under three conditions: 1) degraded auditory stimulus only 2) visual stimulus only, and 3) degraded auditory plus visual stimuli. Under each condition, participants sat in a chair inside a sound attenuating chamber with the door sealed shut. A 50 cm video monitor was placed about 60cm outside the window of the chamber. The monitor was positioned at eye level, about 122cm away from the participant's head. Stimuli were presented using recorded DVDs. For the auditory condition only, the video monitor was turned off, the shade pulled down, and TDH 39-ohm circum-aural headphones were worn. For the

visual condition only, the video monitor was turned on, the shade up, the headphones removed, and the sound turned off.

Stimuli Selection

A set of CVC syllables were used as the stimuli for this study. Each syllable was selected in accordance with the following conditions:

1. Pairs of stimuli were minimal pairs, differing by only one phoneme, the initial consonant
2. All stimuli were accompanied by the vowel /ae/, since it does not involve lip rounding or lip extension.
3. Multiple stimuli were used in each category of articulation, including place (bilabial, alveolar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced)
4. All stimuli were presented without a carrier phrase (citation style)
5. Stimuli were known to elicit McGurk-like responses

Stimuli

For each condition, the same sets of stimuli were administered. The eight stimuli used were as follows:

1. bat
2. cat
3. gat
4. mat
5. pat
6. sat
7. tat
8. zat

The stimuli were presented as either single-syllable stimuli or dual-syllable stimuli. In the single-syllable presentation for the auditory alone or visual alone conditions, only one syllable is presented. In the single-syllable presentation for the auditory + visual conditions, both modalities use the same syllable. In the dual-syllable presentation for the auditory + visual condition, each modality presents a different syllable.

For the auditory + visual condition, half of the stimuli were presented as single-syllable stimuli while the other half used the following dual-syllable stimuli sets:

1. bat-gat
2. gat-bat
3. pat-cat
4. cat-pat

Stimulus Presentation

Audio Signal Degrading: Fourteen talkers provided the speech stimuli for the auditory stimuli. Each talker was recorded through a microphone directly into a computer, using the software program Video Explosion Deluxe. Each talker repeated the set of eight monosyllabic stimuli five times. These auditory files were then run through Praat, a computer program that degraded the speech sound to three combined sine waves: F0 + F1 + F2.

Digital Video Editing: Visual stimuli for the study were obtained by first recording the talkers with a digital video camera; each talker repeated the list of eight stimulus words five times. Stimuli from the recordings were then downloaded and edited using a computer software program, Video Explosion Deluxe. Within this program, auditory stimuli created with the Praat program were dubbed onto the visual representation of a

speech sound. The program was used to create stimuli featuring the same auditory and visual syllables (single-syllable stimuli) as well as stimuli featuring different auditory and visual syllables (dual-syllable stimuli.) The incongruent stimuli were used to analyze McGurk-type integration effects.

Through the computer software program, Sonic MY DVD, stimulus lists were created and burned onto recordable DVDs. Three DVDs were produced for each talker, all with different randomized stimulus orders, to minimize the possibility of effects that can occur from order of stimulus presentation. For this study, the DVDs were played in a DVD player connected to a video monitor.

The testing was done in three presentation conditions: degraded auditory only, visual only, and degraded auditory + visual. Each subject was tested under all three conditions for each talker, with the order of conditions randomized across subjects. For each trial, participants were asked to repeat the word that they thought had been presented. These responses were manually recorded during testing.

Degraded Auditory Alone: Under the degraded auditory alone condition, participants listened to the auditory stimuli of the recorded DVDs through headphones while in the sound attenuating chamber. Randomized orders of DVDs were played for each of the subjects. Participants were seated in a chair in the back of the chamber, facing the video monitor outside the window, though for this condition, the shade in front of the window was pulled down and video monitor turned off in order to remove the visual cues of the talker.

Visual Alone: Under the visual alone condition, participants watched visual stimuli of the recorded DVDs on the video monitor. Randomized orders of the DVDs were played for each of the subjects. Again, participants were seated in a chair within the sound attenuating chamber, facing the video monitor outside the chamber window. They were asked to repeat the syllable that they felt they had perceived. As this condition required the absence of auditory cues, the participant did not wear headphones and the video monitor's sound was turned off.

Degraded Auditory + Visual: As for the previously noted conditions, under the degraded auditory plus visual condition, participants were seated in a sound attenuating chamber, facing the video monitor outside the chamber window. Participants wore a set of headphones in order to listen to the degraded auditory stimulus, and the shade to the chamber window was pulled up to allow viewing of the visual stimulus on the video monitor. A randomized order of DVDs was played for each participant via the video monitor and headphones.

Procedure

Testing Setup

Testing for this study was done at the Ohio State University in a lab room of the Speech and Hearing Department. The lab room, located in a quiet area of the basement, was well lit with fluorescent lighting. Participants were seated in a chair alone the back wall of one of the lab's sound attenuating chambers. All participants sat the same distance away from the chamber's window and the video monitor. Examiner

feedback and subject responses were transmitted through an intercom system in the chamber.

The 50 cm video monitor was place outside the booth approximately 4 feet away from the participant and facing a double-glass window on one wall of the chamber. The video monitor was positioned at eye-level for optimal viewing of the stimuli. The chamber door was completely sealed for all testing to keep the testing area quiet and free of distraction. During the presentation of the degraded auditory alone condition, the shade in front of the window was pulled down and video monitor turned off in order to remove the visual cues of the talker. During the presentation of the visual alone condition, the video monitor's sound was completely turned off. For the auditory and auditory + visual conditions, the participants wore headphones.

Testing Tasks

Each participant was presented with 42 prerecorded DVDs, three videos per talker, each containing a randomized set of 60 stimuli syllables. Each stimulus word was presented multiple times. The syllables presented to the participants consisted of eight stimuli, differing only in the initial consonant. The three DVDs presented for a single talker were randomly shown in each of the following conditions:

1. degraded auditory only
2. visual only
3. degraded auditory + visual

Participants were instructed to verbally respond to what they perceived on the video monitor and/or headphones, while an experimenter transcribed their responses.

Participants were informed that the words they were presented could be both words and nonsense syllables, including phoneme sequences they may or may not encounter in the English language.

Chapter 3: Results and Discussion

Results were analyzed for two types of stimuli; single-syllable stimuli and dual-syllable stimuli. Performance was first assessed for single-syllable stimuli under each presentation condition, which includes degraded auditory only, visual only, and degraded auditory+visual. For single-syllable stimuli, performance was measured in percent correct responses. The degree to which audiovisual integration has occurred can be determined by comparing the number of correct responses for the condition auditory+visual to the number of correct responses for the conditions of auditory or visual only; the greater the improvement for the auditory +visual condition, the more integration has occurred.

Second, performance was assessed for dual-syllable stimuli, which was used for half of the stimuli presentations in the degraded auditory +visual condition. When presenting the dual-syllable stimuli, each modality receives a different syllable. For example, the degraded auditory stimuli may be presented as the syllable “bat” while the visual stimulus is presented as the syllable “gat.” For these stimuli, there is no single “correct” response. Instead, responses are categorized as “visual,” “auditory,” or “other.” Responses categorized as “other” are responses that are different from both the visual and auditory stimuli.

Single-Syllable Stimuli

Figure 1 shows the overall percent correct identification for single-syllable stimuli in the three conditions of auditory, visual, and auditory+visual. The figure indicates that the performance across talkers and subjects for the degraded auditory only and the

visual only conditions was on average about the same. The results also show that listeners were able to integrate the visual and degraded auditory signals to achieve higher performance in the auditory+visual condition.

Figure 2 compares the performance of different talkers in the visual only condition for single-syllable stimuli. The figure shows little variation among talkers, with percent correct identification ranging from 25% (talker DA) to 35% (talker DF.) Figure 3 compares the performance of different talkers in the degraded auditory only condition for single-syllable stimuli. Figure 3 again shows little variation among talkers with percent correct identification ranging from 22% (talker KS) to 37% (talker SS.) Figure 4 compares the performance of different talkers in the auditory+visual condition for single-syllable stimuli. As with Figure 1 and 2, Figure 3, likewise, shows little variation across talkers with percent correct identification ranging from 35% (talker DA) to 52% (talker KS.)

Interestingly, though there is little variation across talkers in Figures 2-4, there is significant variation between talkers when comparing performance across conditions. Figure 5 compares the performance of different talkers in the degraded auditory only and auditory+visual conditions for single-syllable stimuli. The figure shows that talkers who performed well in the degraded auditory only condition, were not necessarily the same talkers who performed well in the auditory+visual condition. Likewise, the talkers who performed poorly in the degraded auditory only condition, were not necessarily the same talkers who performed poorly in the auditory+visual condition. This suggests that talkers who have optimal auditory intelligibility are not necessarily the same talkers that facilitate optimal audiovisual integration. Most notable is the performance of KS, who

had the poorest performance of all talkers in the auditory only condition, but the best performance of all talkers in the auditory+visual condition.

Figure 6 compares the performance of different talkers in the visual only and auditory+visual conditions for single-syllable stimuli. The figure again shows that talkers who performed well in the visual only condition, were not necessarily the same talkers who performed well in the auditory+visual condition. Likewise, the talkers who performed poorly in the visual condition, were not necessarily the same talkers who performed poorly in the auditory+visual condition. In this figure, as in Figure 5, the performance of the talkers in the visual only condition was not indicative of the performance of the talkers in the auditory+visual condition.

Figure 7 compares the performance of different talkers across all three conditions (visual only, degraded auditory only, and auditory+visual) for single-syllable stimuli. This figure clearly shows that talkers who yield the best auditory+visual performance are not necessarily talkers who yield the best auditory or visual performance. Most notable is the difference seen in talkers DA, DF, KD and KS. These talkers are replotted in Figure 8, which specifically compares these four talkers. Talker DA yields the poorest performance auditorily and one of the poorer performances visually, and expectedly, yields the poorest performance when the two modalities are combined for the auditory+visual condition. DF yields the best visual performance and an average auditory performance across talkers, but still yields results very similar to DA for the auditory+visual condition. When compared to the other talkers, the range of KD and KS's degraded auditory only and visual only performances is the greatest with KD's auditory performance exceeding visual performance by 10% and KS's visual

performance exceeding auditory performance by 10%. Both of these talkers produce performances that are some of the lower if not lowest within each condition; however, both talkers yield two of the higher auditory+visual performances.

Figure 9 compares performance of the listeners across all three conditions (visual only, degraded auditory only, and auditory+visual) for single-syllable stimuli. This figure shows that there are substantial differences across listeners as well as across all three conditions. Some listeners showed a good amount of benefit between each single modality and the auditory+visual modality, while others showed very little benefit.

Statistical analysis using a two factor, within subject ANOVA indicated that there was a significant main effect of talker, $F(13, 117)= 3.57$, $p=.008$, $\eta 2=.28$. Follow-up pairwise comparisons did not show any striking pattern of differences. There was also a significant main effect of presentation condition, $F(2,18)= 25.7$, $p<.001$, $\eta 2=.74$. Follow-up pairwise comparisons indicated significant differences between visual only and auditory+visual, and between auditory only and auditory +visual, but not between visual only and auditory only. However, no significant talker by condition interaction was observed.

Pearson r correlations across talkers showed no relationship between the auditory only and visual only performance ($r = -.003$, ns.) Pearson r correlations also showed no relationship between visual only and auditory+visual performance ($r = .34$, ns), or between auditory only and auditory+visual performance ($r = .30$, ns.

Dual-Syllable Stimuli

Figure 10 shows the overall percent of response types for dual-syllable stimuli. This figure shows that when talkers were presented with dual-syllable stimuli where the syllables for each modality are incongruent, talkers chose the visual syllable presented more often than the degraded auditory syllable presented. However, the majority of the responses were something other than the either of the two syllables presented through the two modalities.

Figure 11 shows the performance of different talkers in their production of fusion McGurk responses for dual-syllable stimuli. This figure shows that there is some variance across talkers in number of fusion McGurk responses they elicited. Performance across talkers ranged from 12% (talker KD) to 30% (talker DA.)

Figure 12 shows integration performance for dual-syllable stimuli (fusion McGurk responses) for each talker. Again, substantial differences across talkers are seen in the amount of integration produced. Also plotted here is the percent audiovisual improvement for the single-syllable condition. As can be seen, the differences in integration performance for dual-syllable stimuli are not predicted by the degree of audiovisual integration in the single-syllable conditions. For example, talkers DA and LG show the two lowest percents of improvement for the auditory+visual condition, but show the two highest percents for production of fusion McGurks. On the opposite end of the spectrum, KS shows, by far, the highest percent improvement for the auditory+visual condition, but surprisingly, an average production of fusion McGurks.

Pearson r correlations across talkers showed no relationship between the percent correct auditory performance in the single-syllable presentation and the percent

auditory responses in the dual-syllable presentation (r = .15, ns.) Pearson r correlations across talkers also showed no relationship between the percent correct auditory performance in the single-syllable presentation and fusion McGurks (r = .15, ns), or between the percent correct visual performance in the single-syllable presentation and fusion McGurks (r = .18, ns.)

Overall, these data suggest that talker differences are a crucial variable in audiovisual integration; however, integration performance cannot be predicted by overall differences in single-modality performance for individual talkers.

Chapter 4: Summary and Conclusion

Taken together, results show that there are differences across talkers that can account for the variability in listeners' ability to integrate speech. This can be seen by the fact that Pearson correlation coefficients showed no significant relationships across talkers among any of the presentation conditions. This suggests that audiovisual integration cannot be predicted from performance in either single modality. In addition, the differences in integration performance for dual-syllable stimuli are not predicted by the degree of audiovisual integration in the single-syllable conditions.

Results also showed that by degrading the speech signal to three sine wave speech, talkers' ability to elicit strong auditory performance was substantially lowered. This is seen in the auditory condition's overall percent correct score, which was 33%.

The results from the present study are only a preliminary look into how talker differences affect speech perception. The present study only examined the benefit of the auditory and visual modalities across talkers. Future work should look more closely at specific talker differences, such as talkers' measures of visible articulation. This includes lip opening, lip rounding, jaw movement, etc., as well as the talkers' measures of the acoustic speech tokens including formant values and transitions, temporal measures, etc. Through this examination it can be evaluated whether particular aspects of individual articulation are correlated with improvements in audiovisual integration, thus moving toward a comprehensive depiction of the individual talker characteristics of a "good" talker for situations requiring audiovisual integration of the speech signal.

The results of the present study have long-term implications for the development of aural rehabilitation programs for individuals with hearing impairments. Through finding the characteristics of talkers that lead to optimal audiovisual integration, speech pathologists and audiologists can produce effective training tools for patients.

References

Chen, F. (1980). Acoustic characteristics of clear and conversational speech at the segmental level. Cambridge, MA: Massachusetts Institute of Technology.

Clark, C. (2005). *Effects of Long Term Audio-Visual Versus Audio-only Experience on Multimodal Speech Perception.* Senior Honors Thesis. The Ohio State University.

Gagne, J.P., Rochette, A-J., and Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Communication, 37,* 213-230

Grant, K.W., and Braida, L.D. (1991). Evaluating the Articulation Index for audiovisual input. *Journal of the Acoustical Society of America 89*, 2952-2960.

Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences, *Journal of the Acoustical Society of America, 104*, 2438-2450.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90 (5),* 99-114.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

Munhall, K.G., Kroos, C., Jozan, C., and Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics 66*, 574-583.

Payton, K.L., Uchanski, R.M., and Braida, L.D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America* 95, 1581-1592.

Picheny, M.A., Durlach, N.I. and Braida, L.D. (1985). Speaking clearly for the hard of

hearing I; Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research 28*, 96-103.

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science 212*, 947-950.

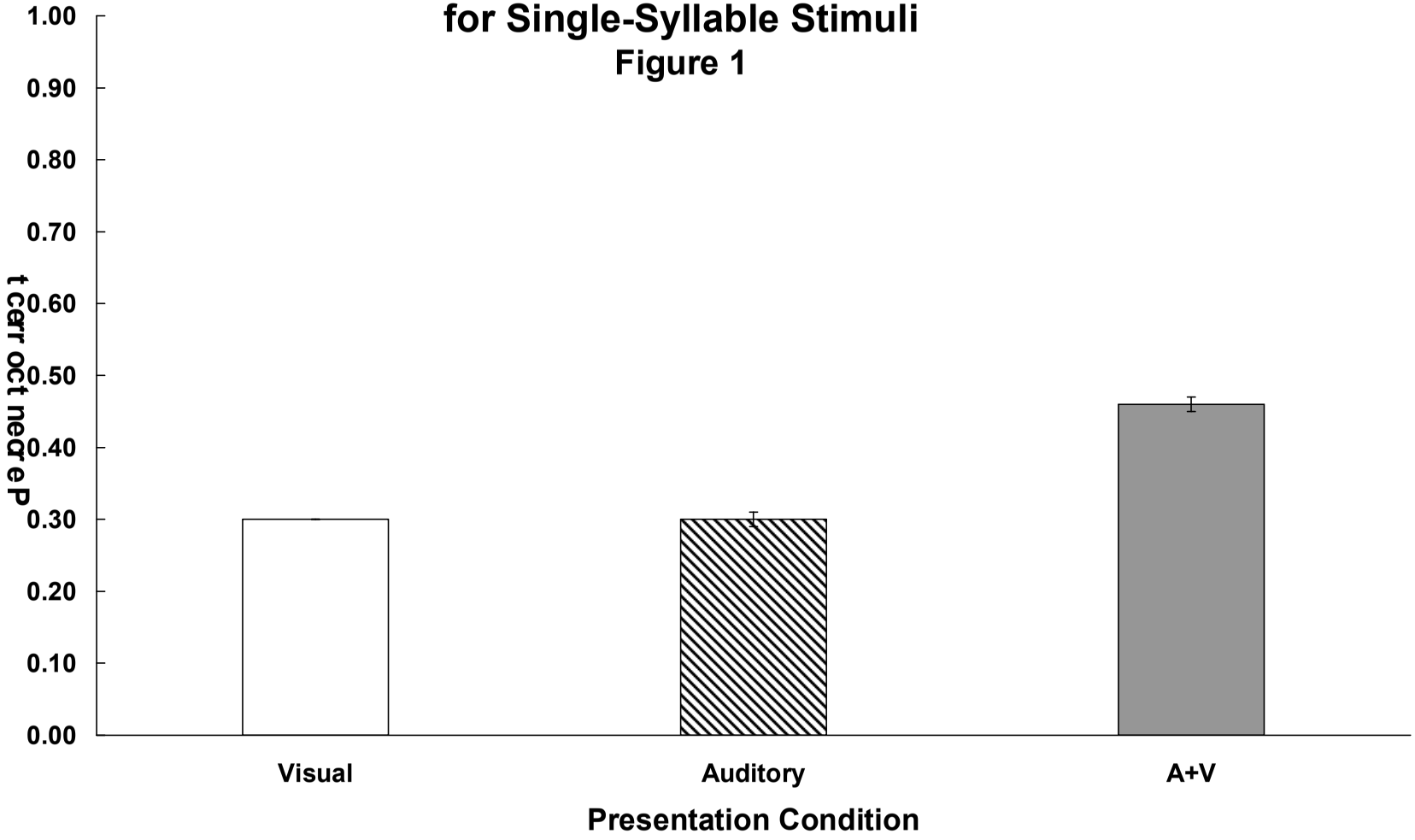
Sumby, W.H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America 26*, 212-215.

Uchanski, R.M., Miller, K.M., Reed, C.M., and Braida, L.D. (1992). Effects of token variability on vowel identification. In Schouten. E.H. (Ed.), *The auditory processing of speech*. New York: Mouton de Gruyter.

List of Figures

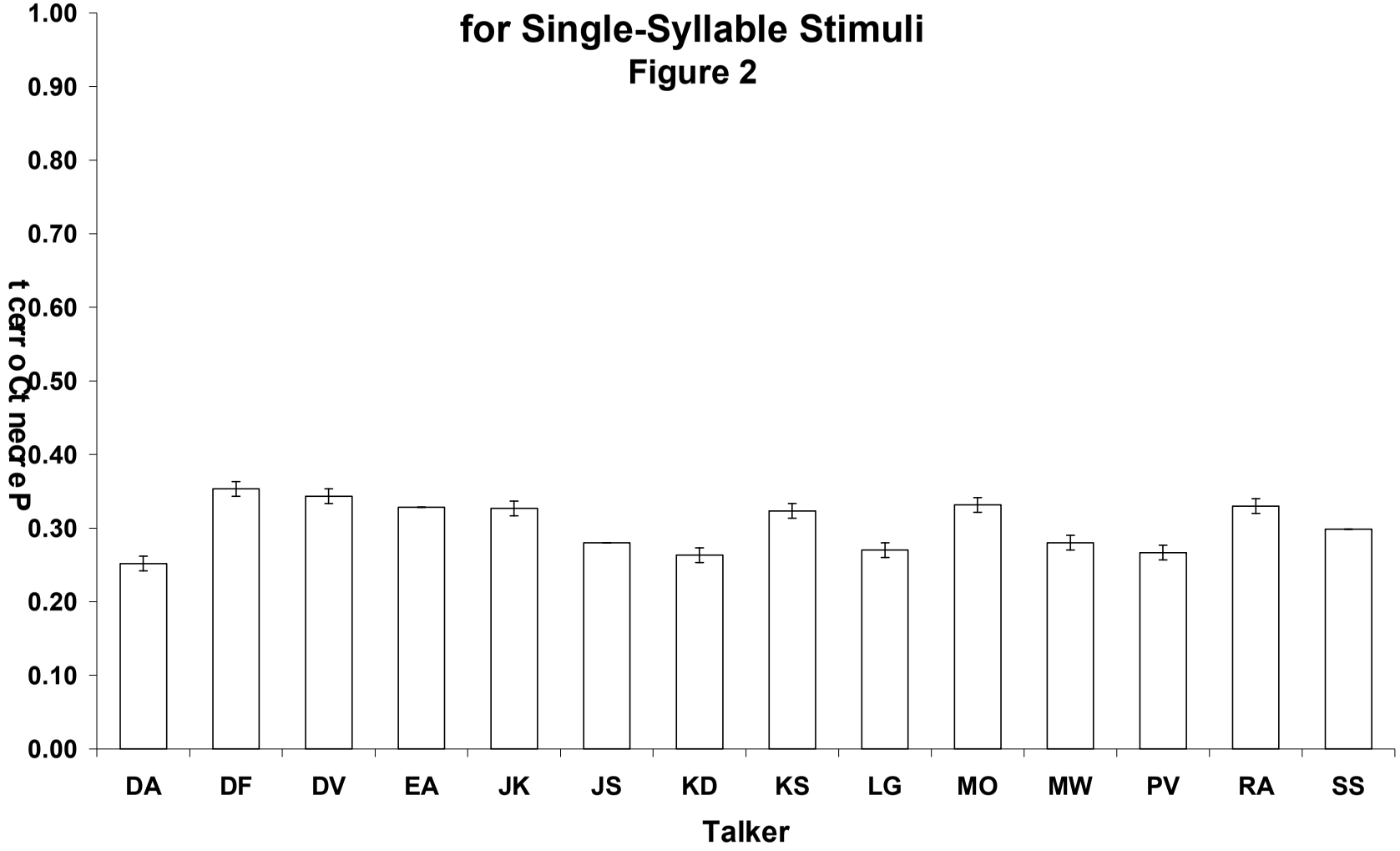
Overall Percent Correct by Presentation Condition for Single-Syllable Stimuli

Figure 1



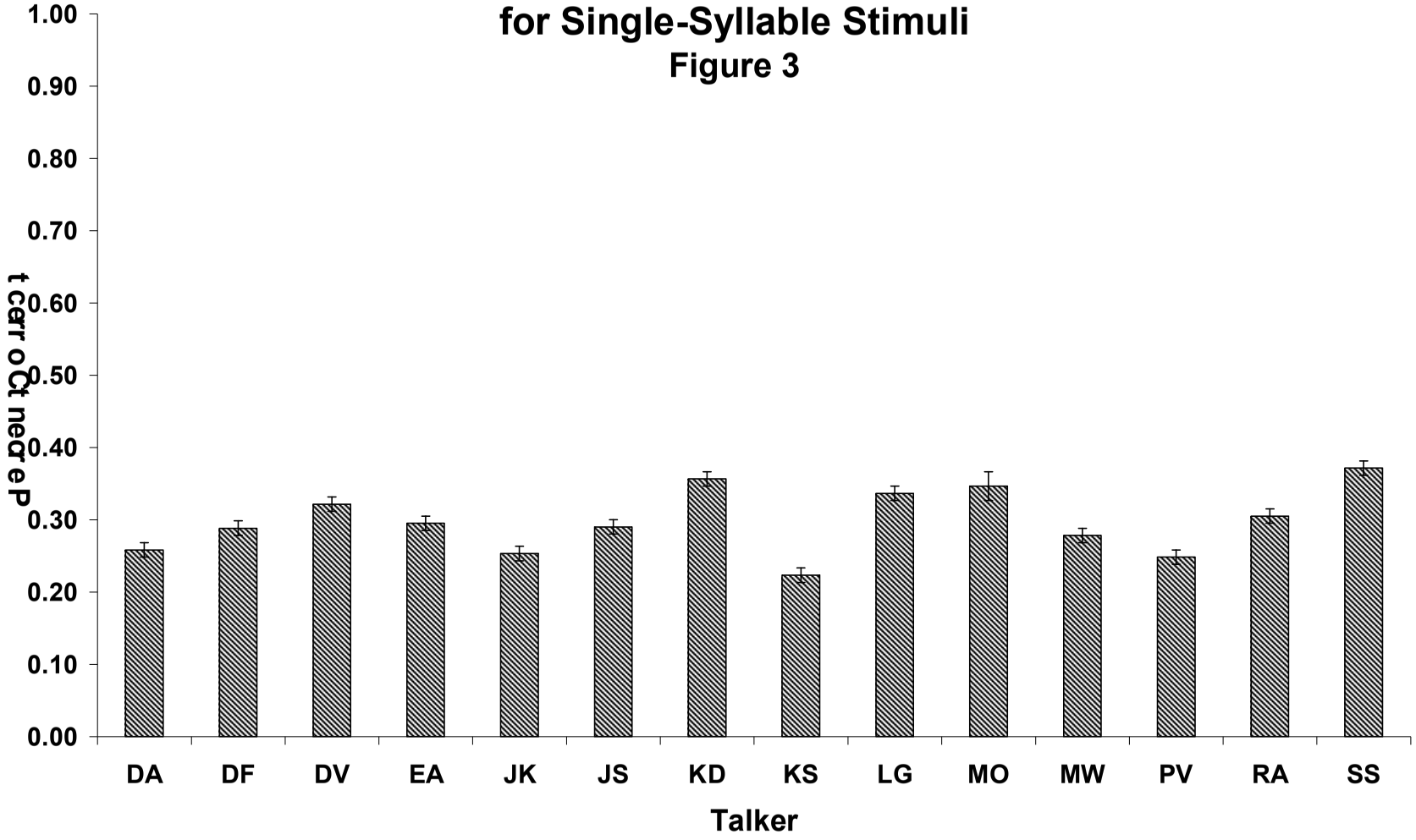
Percent Correct Identification in Visual Conditions for Single-Syllable Stimuli

Figure 2



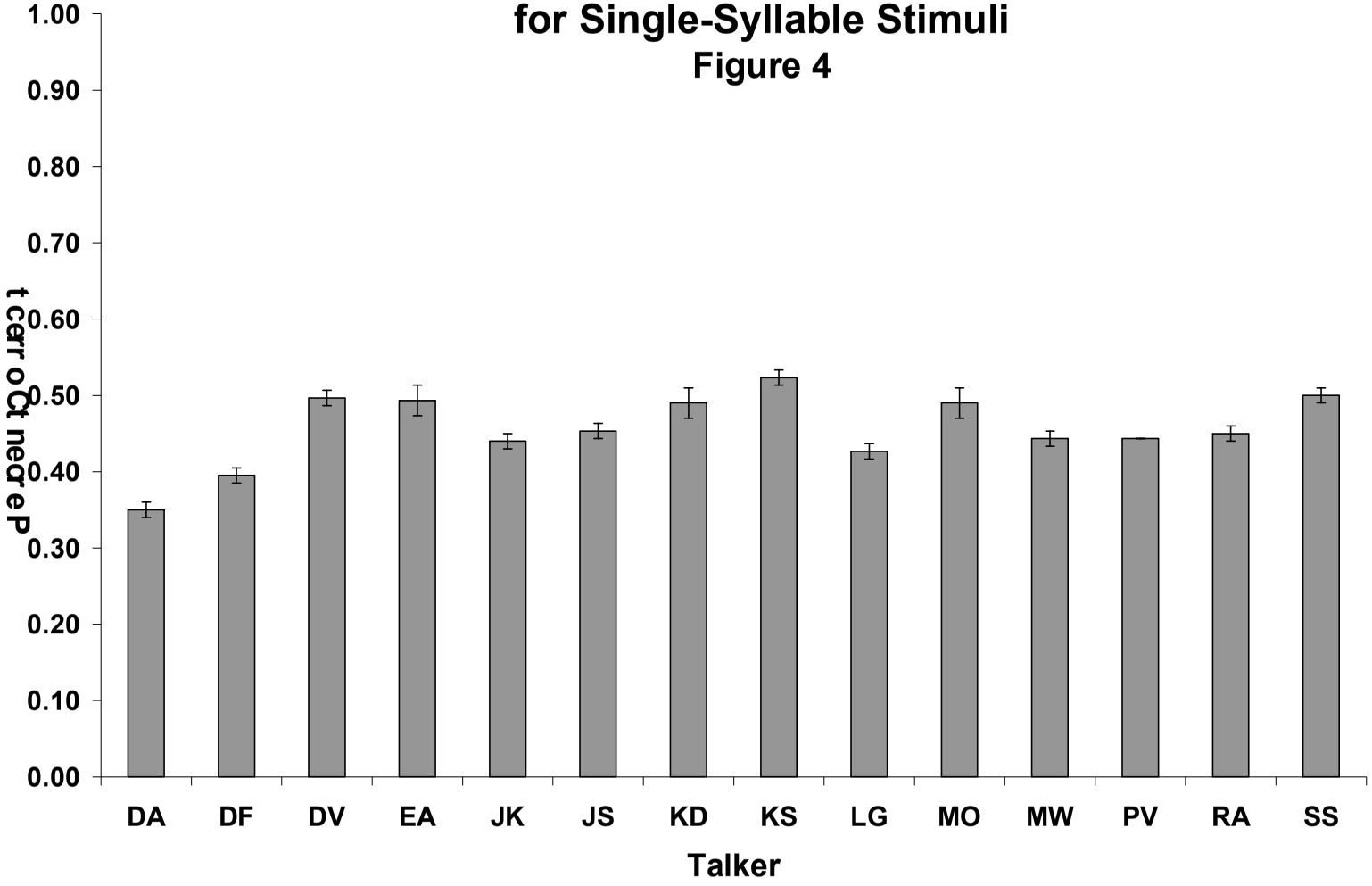
Percent Correct Identification in Auditory Conditions for Single-Syllable Stimuli

Figure 3



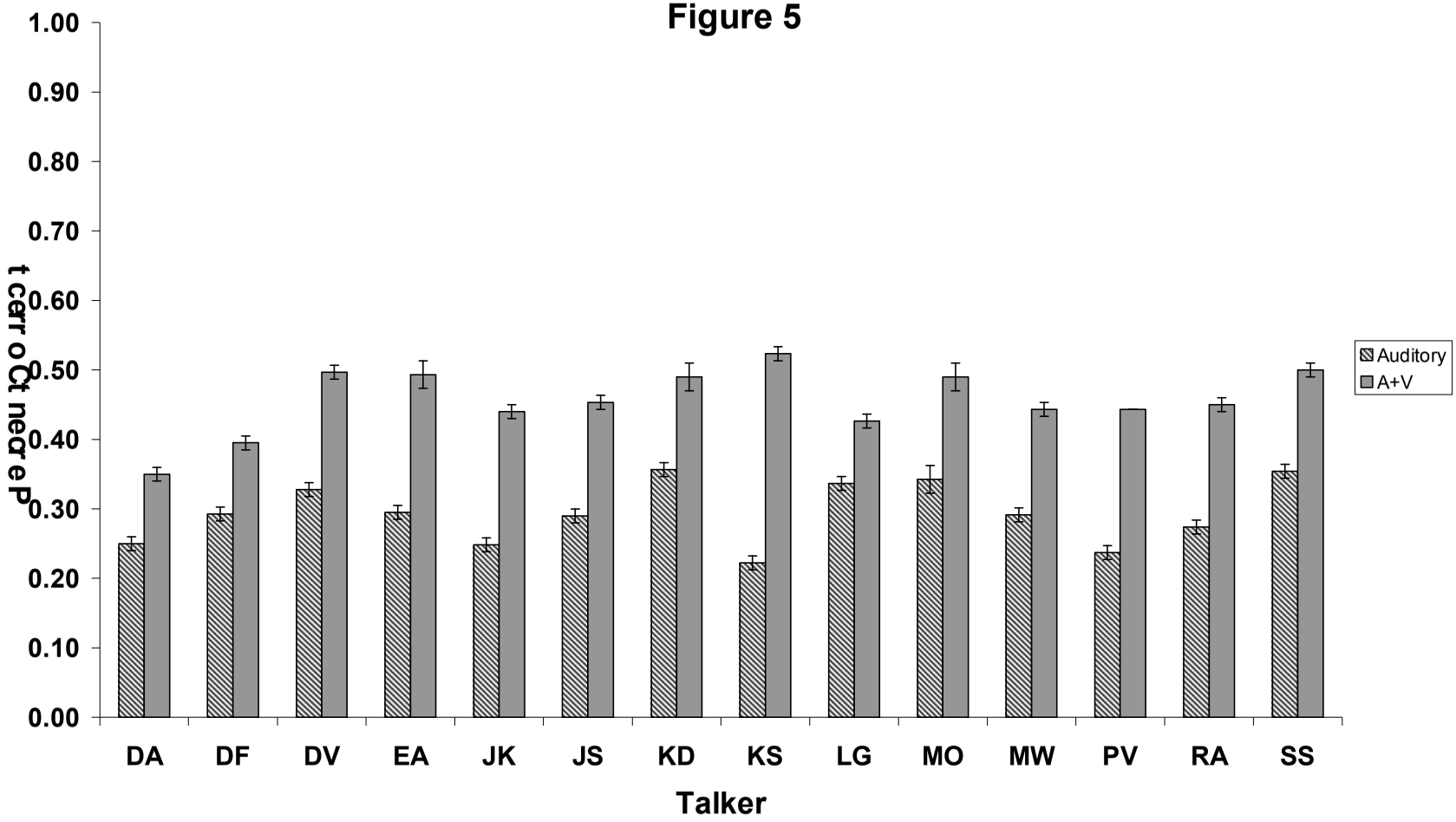
Percent Correct Identification in A+V Conditions for Single-Syllable Stimuli

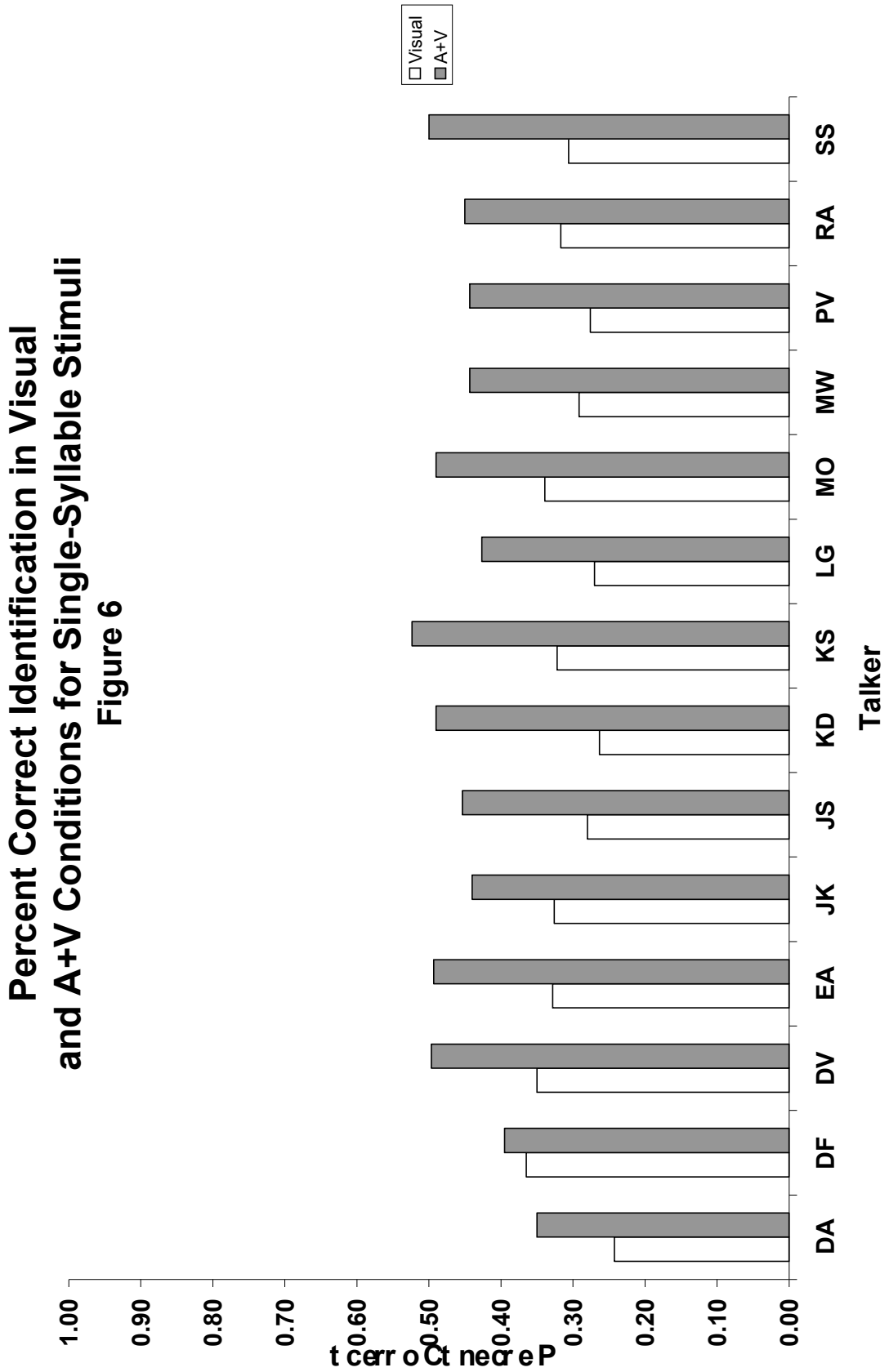
Figure 4



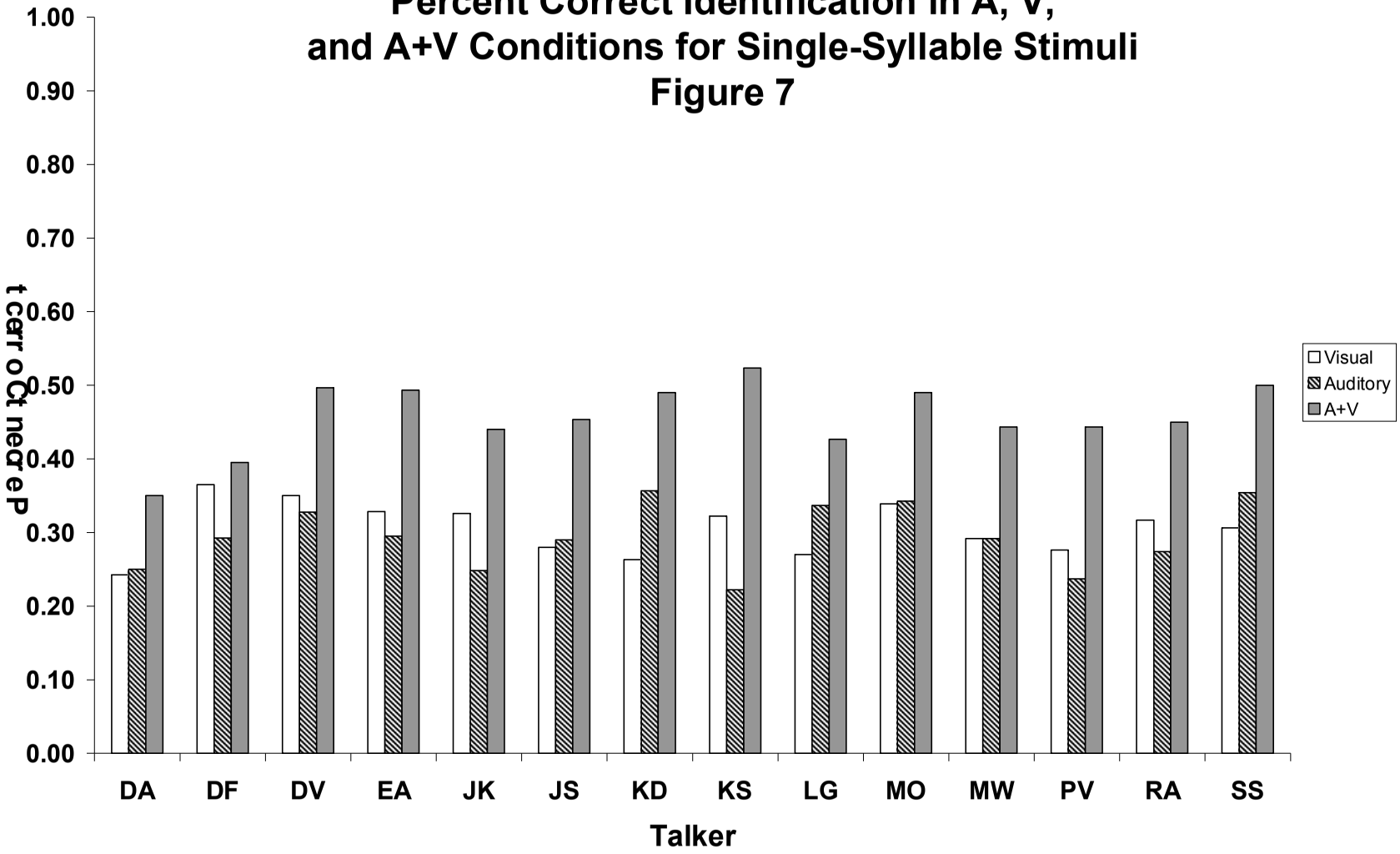
Percent Correct Identification in Auditory and A+V Conditions for Single-Syllable Stimuli

Figure 5



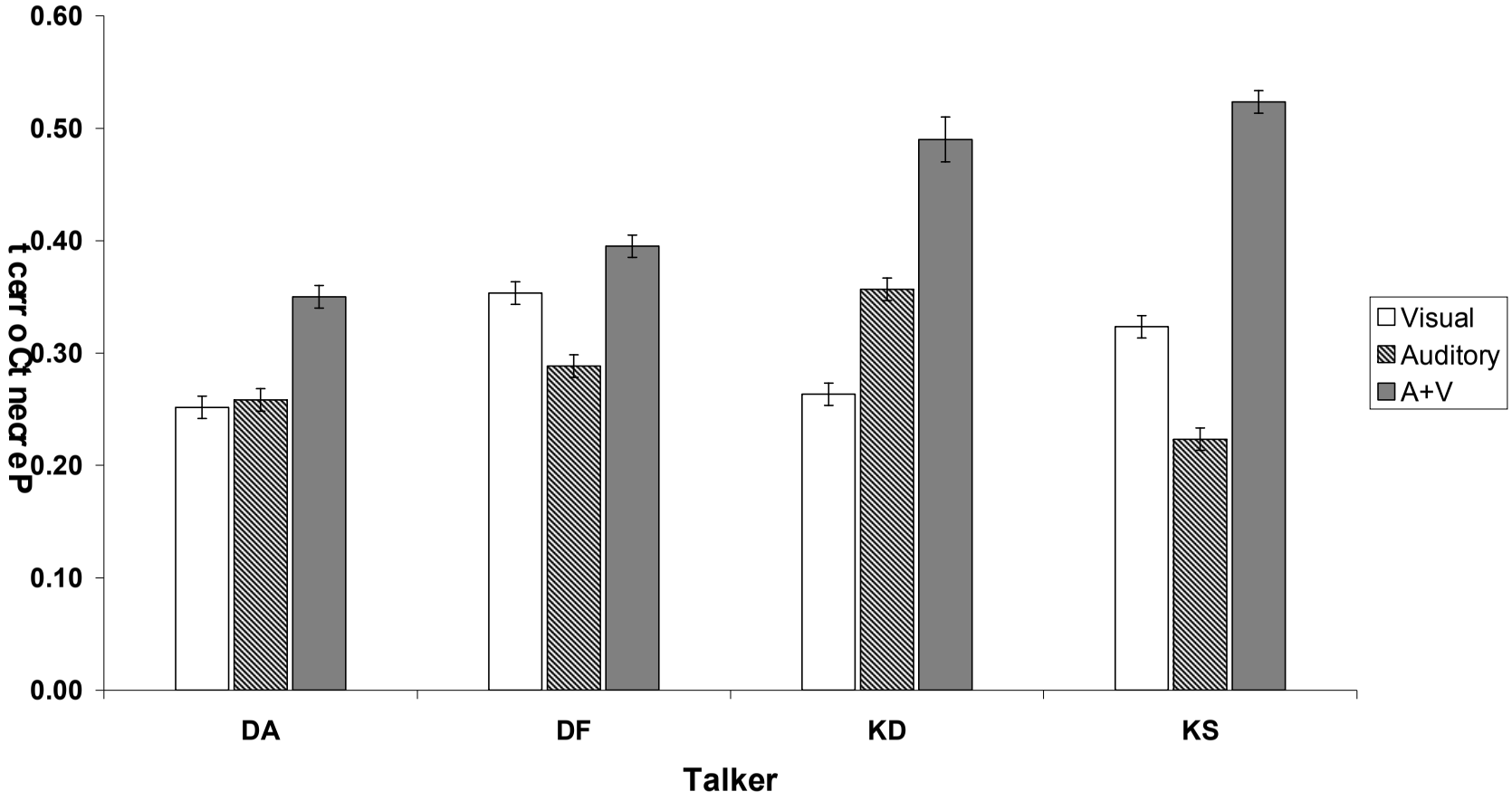

Percent Correct Identification in Visual and A+V Conditions for Single-Syllable Stimuli
Figure 6
Visual
A+V
Percent Correct
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
DA
DF
DV
EA
JK
JS
KD
KS
LG
MO
MW
PV
RA
SS
Talker

Percent Correct Identification in A, V, and A+V Conditions for Single-Syllable Stimuli
Figure 7



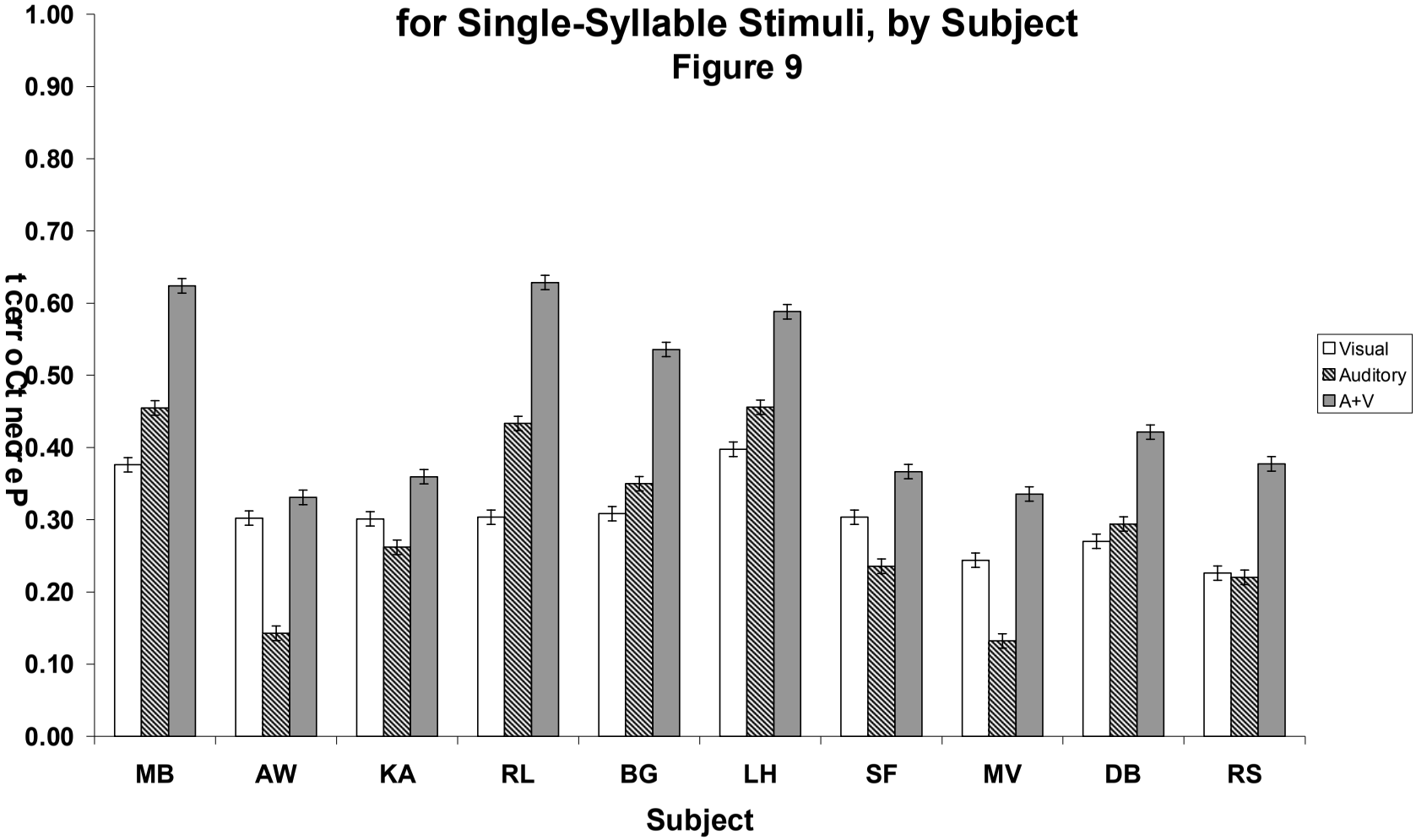
Percent Correct Identification in V, A, and A+V Conditions for Select Talkers

Figure 8



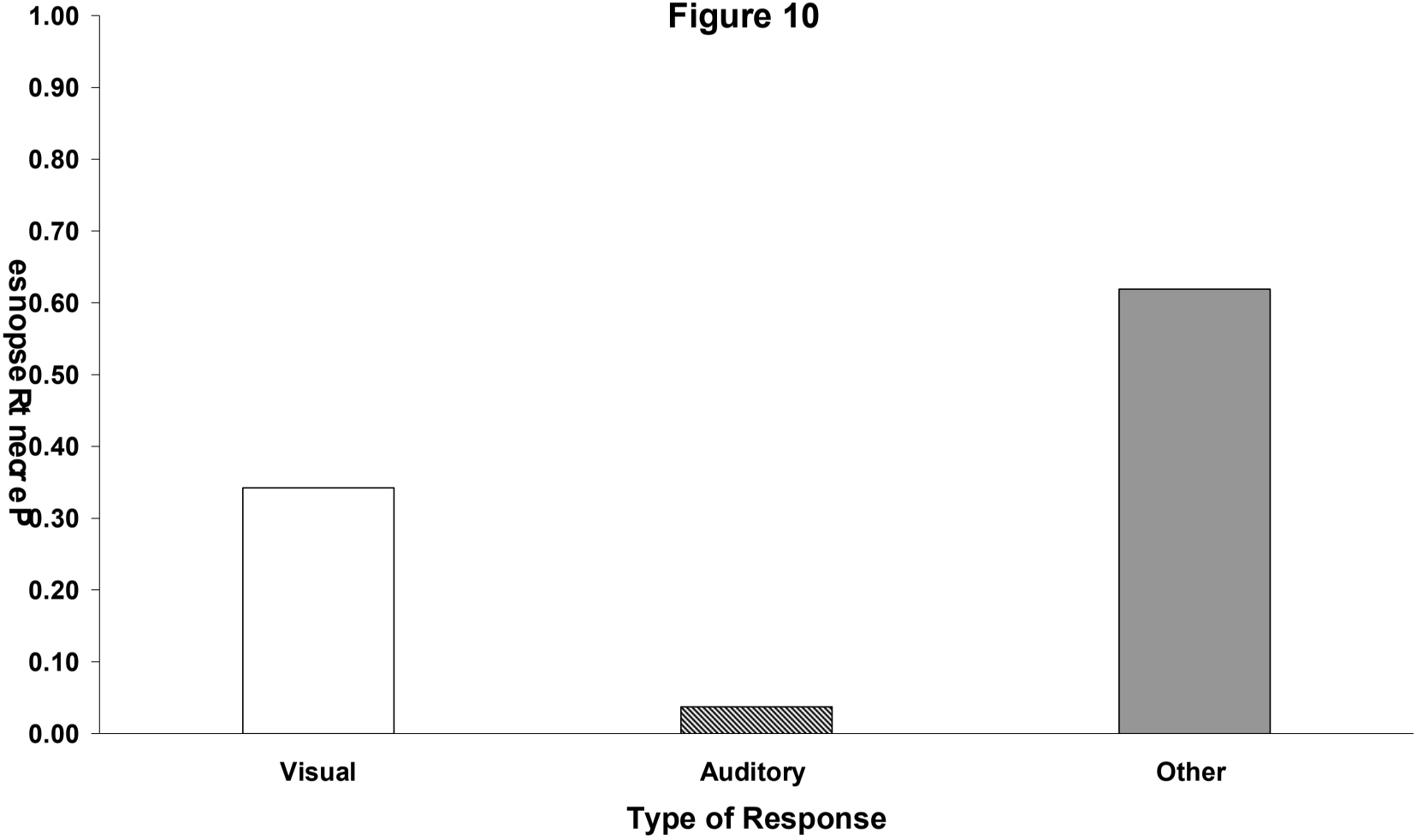
Percent Correct Identification in V, A, and A+V Conditions for Single-Syllable Stimuli, by Subject

Figure 9



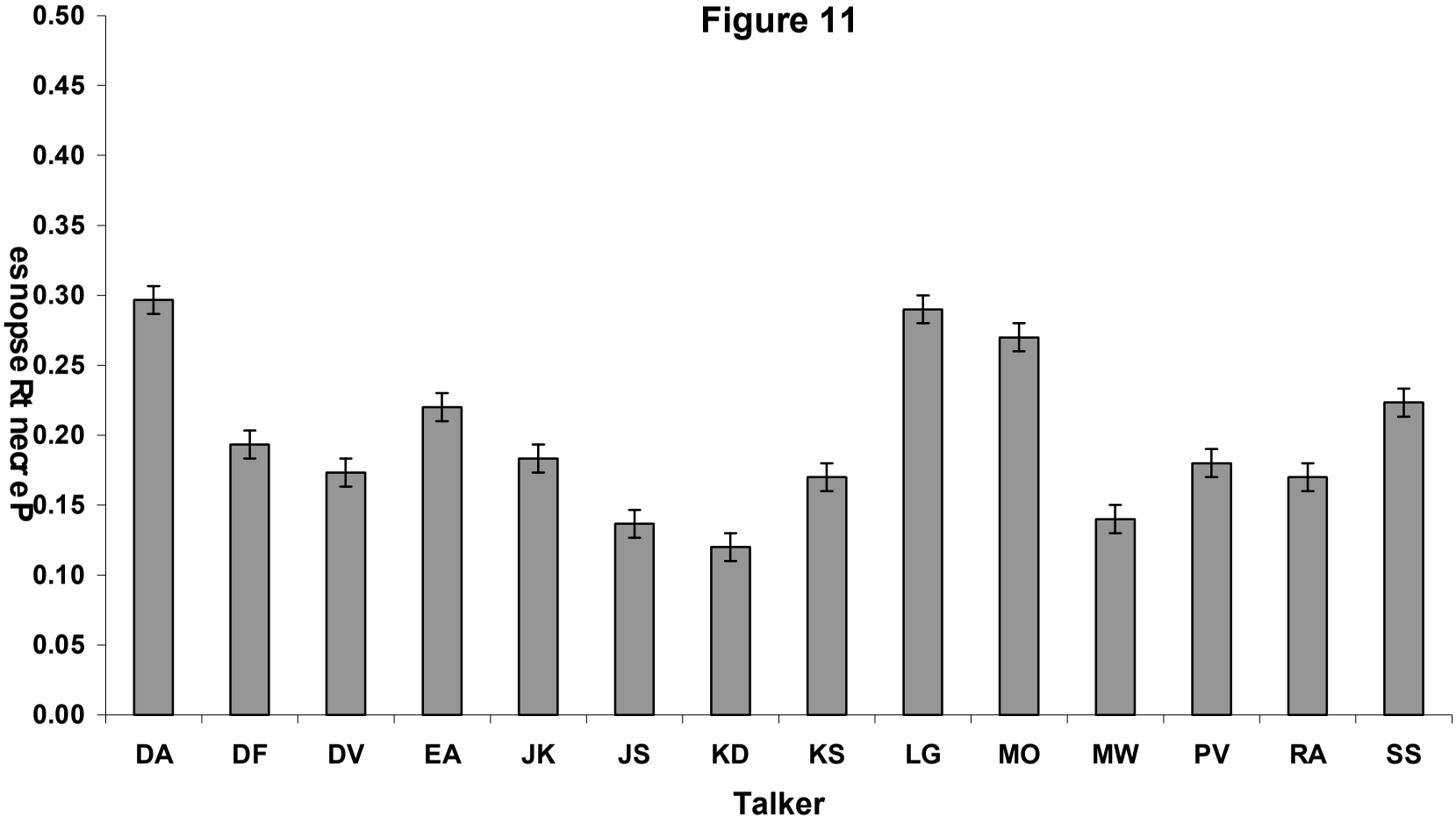
Percentage of Responses for Dual-Syllable Stimuli

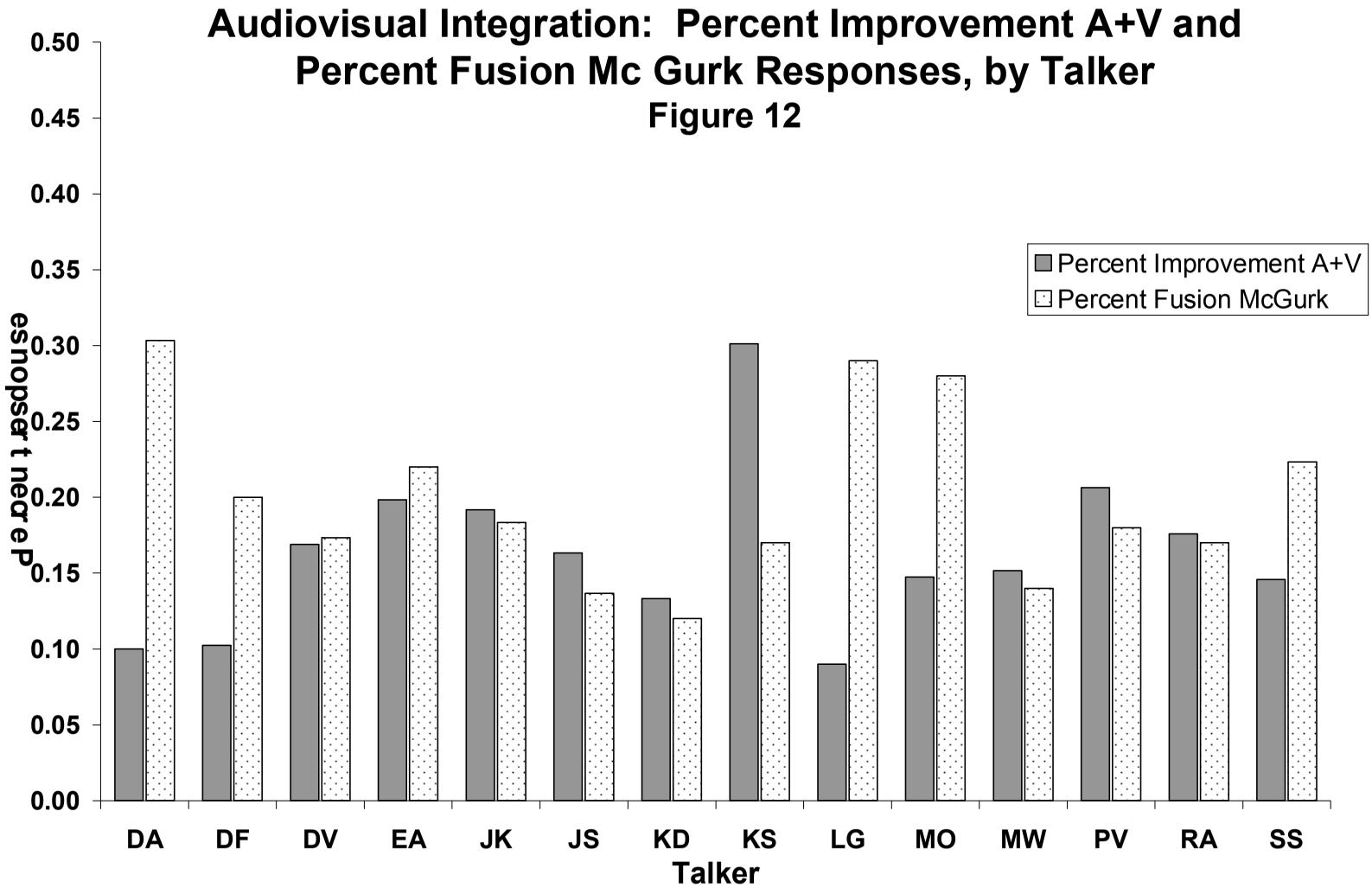
Figure 10



Percentage of Fusion McGurk Responses for Dual-syllable Stimuli

Figure 11




Audiovisual Integration: Percent Improvement A+V and Percent Fusion Mc Gurk Responses, by Talker
Figure 12
Percent Improvement A+V
Percent Fusion McGurk
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
DA
DF
DV
EA
JK
JS
KD
KS
LG
MO
MW
PV
RA
SS
Talker